Handling Missing and Unreliable Information in Speech Recognition

Phil Green, Jon Barker, Martin Cooke, and Ljubomir Josifovski
{p.green, j.barker, m.cooke, l.josifovski}@dcs.shef.ac.uk
Department of Computer Science, University of Sheffield
Sheffield S1 4DP, UK

Abstract

In this work, techniques for classification with missing or unreliable data are applied to the problem of noise-robustness in Automatic Speech Recognition (ASR). The primary advantage of this viewpoint is that it makes minimal assumptions about any noise background. As motivation, we review evidence that the auditory system is capable of dealing with incomplete data and, indeed, does so in normal listening conditions. We formulate the unreliable classification problem and show how it can be expressed in the framework of Continuous Density Hidden Markov Models for statistical ASR. We describe experiments on connected digit recognition in noise in which encouraging results are obtained. Results are improved by 'softening' the missing data decision. We argue that if the noise background is unpredictable it is necessary to integrate primitive processes which identify coherent spectral-temporal regions likely to be dominated by a single source with a generalised recognition decode which searches for the best sub-set of regions which match a speech source. We describe an implementation of a multi-source decoder using missing data recognition and show how it improves recognition results for non-stationary noises.

1 Sensory Occlusion in the Auditory System

The 'missing data' problem in computer vision, caused by occlusion, has been the subject of a number of studies e.g. (Ahmad and Tresp, 1993; Ghahramani and Jordan, 1994). The equivalent problem in audition has received far less attention because it is counter to intuition: while objects in a visual scene are predominantly opaque, acoustic signals combine additively. Consequently, techniques for robust automatic speech recognition (ASR) typically aim for near-perfect allocation of the acoustic mixture into additive contributions from constituent sources (see Furui (1997) for a review), making use of noise source models.

This paper takes as its starting point the alternative hypothesis - that incomplete data is a valid characterisation of the normal listening situation. Our supporting argument is, in summary:

Listeners can cope with missing data. Natural speech signals which have undergone deliberate spectro-temporal excisions typically show remarkably little decrease in intelligibility (Strange et al., 1983; Steeneken, 1992; Warren et al., 1995; Lippmann, 1996). There are counterparts to these experiments in everyday listening e.g. interfering signals, band-restricted transmission, channel noise over telephone lines, hearing disorders.

Masked data is effectively missing data. The neural code for signal detection exhibits what has been called the 'capture effect' (Moore, 1997), in that locally more intense sound components dominate the neural response. Weaker sound components do not contribute to the neuronal output: they are masked and therefore can be considered missing for the purposes of further processing.

The auditory nervous system handles simultaneous signals. The auditory system must do more than process isolated speech sources. In Bregman's terms (Bregman, 1990), the 'auditory scene analysis' problem involves organising multiple, concurrent signals into associated perceptual streams. However, the evidence for each stream will frequently be incomplete.

Compressive acoustic representations make disjoint allocation a good approximation. Due to the large dynamic range of speech signals and the compressive representations that this dynamic range demands, most spectral-temporal regions are dominated by one source, and in regions where one source dominates the noisy representation is very close to the clean representation of that source.

Application of missing data techniques in robust ASR requires a solution to two problems: (i) identification of reliable spectro-temporal evidence; and (ii) modification of recognition algorithms to handle incomplete data. In the next section we address (ii).

2 Classification with Unreliable Data

The classification problem in general is to assign an observation vector x to a class C. In the missing data case, a preceding process has partitioned x into reliable and unreliable parts, (x_r, x_u). We consider the case in which the 'true' values of the unreliable data can be confined by *bounds* : if x is a spectral energy vector in which the unreliable channels are contaminated by additive noise, the speech energy in these channels must lie between 0 and the observed value x_u . The likelihood $f(x|C)$ cannot be evaluated in the normal manner. There are two alternatives:

data imputation: estimate values for the unreliable components of x, producing a complete observation vector estimate $\hat{x}$, and to then proceed with classification using $f(\hat{x}|C)$.

marginalisation: base classification solely on reliable components, effectively integrating over the unreliable components. In the case of complete ignorance of the unreliable data classification is based on $f(x_r|C)$.

In this paper we concentrate on marginalisation. See (Josifovski, 1999) for companion work and (Raj et al., 2000) for other work on data imputation.

3 Application to Robust Automatic Speech Recognition

In conventional Continuous Density Hidden Markov Model Speech Recognition, each chosen speech unit is represented by a trained HMM with a number of states. The states correspond to the classes of the last section. Each state is characterised by a multivariate mixture Gaussian distribution over the components of x, from an observation sequence X. The parameters of these distributions, together with state transition probabilities within models, are estimated in an EM fashion, commonly using the Baum-Welch algorithm. In our work, the models are trained on clean data: there is no re-training for noise conditions. A decoder (usually implementing the Viterbi algorithm) finds the state sequence having the highest probability of generating X. We show in (Cooke et al., 1999) that in these conditions the *'bounded- marginal'* estimation of $f(x|C_i)$ can be written as

$$f'(x|C_i) = \sum_{k=1}^{M} P(k|C_i) f(x_r|k, C_i) \int f(x_u|k, C_i) dx_u$$

Where the $P(k|C_i)$ are the M mixture coefficients for the distribution associated with C_i .

The first term in this equation is the marginal distribution over the reliable vector components. The integral term introduces constraints on the true values of the unreliable components. In the complete ignorance case it reduces to 1. In the bounded case it represents counter-evidence against the hypothesis of class C_i. For multivariate Gaussians, the integral required to evaluate the bounded marginal can be approximated by a difference of error functions.

4 Speech Recognition Experiments

In the experiments reported here, the task is connected digit recognition with controlled amounts of added noise from various noise sources. This is a popular testbed for robust ASR (Pearce and Hirsch, 2000). The TIDigits corpus of digit sequences was used. Acoustic vectors were obtained via a 32 channel auditory filter bank (Cooke, 1991) with centre frequencies spaced linearly in ERB-rate from 50 to 8000 Hz. The instantaneous Hilbert envelope at the output of each filter was smoothed with a first order filter with an 8 ms time constant, and sampled at a frame-rate of 10 ms (this is the same representation as employed in (Cooke et al., 1999) except here 32 channels are being used rather than 64). Finally, a cube root compression was applied to the frame of energy values. HTK (Young and Woodland, 1993) was used for training, and an in-house C++ decoder for recognition. Twelve models ('1'-'9', 'oh', 'zero' and 'silence') consisting of 8 no-skip, straight-through states with observations modeled with a 10 component diagonal Gaussian mixture were trained on clean speech. An additional 1-state silence model was used to model the brief inter-digit pauses that may occur during long digit strings. Factory noise and Lynx helicopter noise from the NOISEX database (Varga et al., 1992) were added (with random start points) at SNRs from +20dB to 0dB to a subset of the TIDigits test set consisting of 240 digit strings. For bounded marginalisation, the lower bound was set to 0 and the upper bound to the value of the noisy speech mixture at each time-frequency point.

Results are shown in Figure 1 (Factory Noise) and Figure 2 (Lynx Helicopter Noise). In these figures,

- *'Raw Filter Bank'* shows how performance with models trained on clean data deteriorates rapidly when these models are used for recognition in noise.
- The *'A Priori'* curve uses knowledge of clean

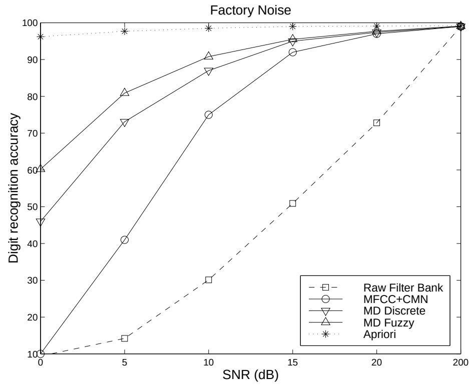

Figure 1: Recognition results for Factory Noise.

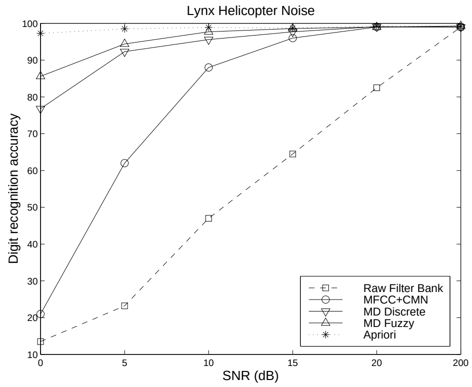

Figure 2: Recognition results for Lynx Helicopter Noise.

speech and noise prior to mixing to find the true local SNR. A time/frequency 'pixel' is judged to be reliable when its local SNR higher than some fixed threshold, typically 7 dB. This defines a missing data mask, illustrated in Figure 3. Using knowledge of the clean speech and the noise is of course cheating but indicates the high performance upper-bound, and provides a dramatic proof of concept - recognition rates hold up to a near-human level of performance even when very few points pass through the mask.

- *'MFCC+CMN'* (Mel Frequency Cepstral Coefficients and Cepstral Mean Normalisation) is a standard technique for improving ASR robustness , for comparison.

For the '*MD Discrete*' results, an estimate of the local SNR is obtained by using the first 10 frames (preceding the speech) to derive a noise estimate for each frequency band. The noise is assumed to be statistically stationary. We have experimented with more advanced noise estimators (Vizinho et al., 1999). A mask obtained by this technique is compared to an a priori mask in Figure 3. Results for noise estimate masks are superior to the standard techniques and, for Lynx helicopter noise, come encouragingly close to that for a Priori Masks, except in severe noise conditions. Results are worse for factory noise, which contains some unpredictable components (e.g. hammer blows). We return to this point in section 6.

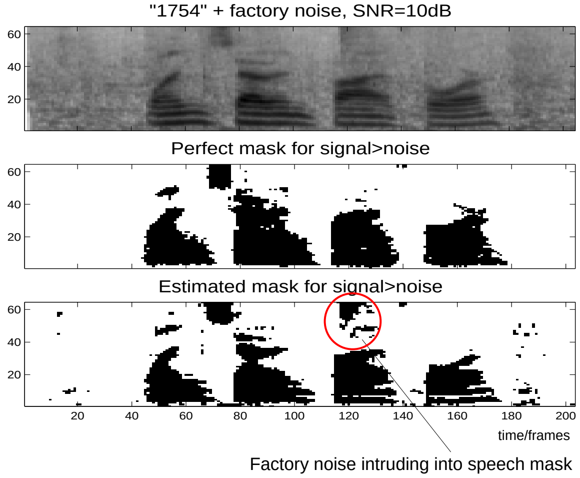

Figure 3: Comparison of the 'a priori' mask and the SNR mask.

5 Using soft decisions for separation

In realistic conditions, we will only be able to obtain an imperfect statistical estimate of the noise, and for any individual pixel the real noise will in any case differ from the mean estimate. If we accept error in our noise estimate, we accept error into the judgement of whether it is the speech or the noise that dominates a particular spectro-temporal region. These errors are made concrete and irreversible when using a discrete mask.

We therefore 'soften' the missing data mask (Barker et al., 2000b). Rather than use either 0 or 1, we use a continuous value in the range [0.0, 1.0] which is interpreted in the missing data probability calculation as "the probability that the point is dominated by the speech signal".

For missing data with **discrete** masks, each components of the feature vector, x, is first classified as either reliable or unreliable (i.e. masked by the noise). The contribution each feature makes to the likelihood of the observation, $f(x|C)$, will depend on how that feature is classified. Assuming an $f(x|C)$ where the components of x are independent:

$$\overline{f(x|C)} = \prod_{i\in r} f_i(x_i|C) \prod_{j\in u} \frac{1}{x_j} \int_0^{x_j} f_j(x_j|C)dx_j$$

With a mask containing **soft decisions** the probability due to each feature vector component becomes a weighted sum of the reliable and unreliable probability terms:

$$\overline{f(x|C)} = \prod_{i=1}^{N} \left(w_i f_i(x_i|k) + (1-w_i)\frac{1}{x_i}\int_0^{x_i} f_i(x_i|k)dx_i\right)$$

As spectral features are, in practice, not independent, following section 3, we extend the above to employ a Gaussian mixture model in which the full distribution $f(x|C)$ is composed of a weighted sum of Gaussian distributions in which the features are independent.

$$f(x|C) = \sum_{k=1}^{M} P(k|C) f(x|C,k)$$

5.1 Generating fuzzy masks

In the current work we employ a simple stationary noise estimate for all noise types. For non-stationary noises the error in the estimate is likely to have a non-Gaussian distribution. Accepting this, we have not attempted to estimate ideal fuzzy masks, but have instead generated a mask of values between 0 and 1 by compressing with a simple sigmoid function (illustrated in Figure 3) with empirically derived parameters. The mapping is of the form:

$$f(x) = \frac{1}{1+e^{-\alpha(x-\beta)}}$$

where α is the sigmoid slope, and β is the sigmoid center. Appropriate values for these parameters are found via a series of tuning experiments. For large values of α the sigmoid becomes steep and the resultant fuzzy mask approximates a discrete missing data mask. In this case we are implicitly assuming a small variance in the noise estimation error. At the other extreme, as the value of α tends to 0, we approach a mask where all values are 0.5. If $\alpha = 0$, we are assuming no knowledge of the noise and admitting maximum uncertainty into the mask.

Our use of fuzzy masks has parallels with the work of Renevey and Drygajlo in which a fuzzy mask is used in conjunction with missing data imputation (Renevey and Drygajlo, 2000).

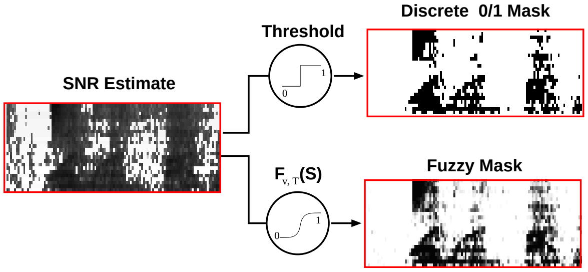

Figure 4: The difference between discrete and fuzzy missing data masks.

5.2 Experiments with fuzzy masks

Within the same experimental framework as section 4, a preliminary series of tuning experiments was run to find appropriate values for the parameters and of the sigmoid function employed in the fuzzy mask technique. Informal tests were conducted to find a suitable range of values, and then tests with a set of 240 utterances were run over a grid of α and β values in these ranges. These tuning experiments were run using both the Lynx helicopter noise and the factory noise corrupted data, and at a range of SNRs. It was found that the optimal α and β values were largely independent of SNR but were dependent on noise type. This point is discussed further below.

The curves labelled *'MD fuzzy'* in Figure 1 and Figure 2 show results for factory and Lynx helicopter noise at the range of SNRs tested. Consider first the factory noise results, (Figure 1). Moving from a discrete mask to a fuzzy mask leads to further performance improvement with the largest gains made at the lower SNRs (the 0 dB result increases from 46% to 60%). The fuzzy mask results here use a sigmoid that has been tuned using the factory noise corrupted data.

For Lynx helicopter noise, we see a roughly similar pattern of improvements though starting from a higher baseline. The Lynx helicopter noise is more stationary and hence the baseline missing data system performs better. The result of 85% accuracy at 0dB, without model re-training, is particularly noteworthy.

Although fuzzy masks ameliorate the problems of discrete masks and poor noise estimates, they are not the full solution to the robust recognition problem. Fuzzy masks soften noise estimation errors but ideally we would like the mask construction to be informed by top-down knowledge from the speech models themselves. This is the topic of the next section.

6 Decoding Multiple Sources

In contrast to conventional decoding, where all the observations are assumed to belong to the source being recognized, the task of a multi-source (MS) decoder (Barker et al., 2000a) is to determine the most likely model state sequence at the same time as deciding which observations to use, and which to ignore as ‘background’.

We assume that we have models for the speech source, but in contrast with approaches such as Parallel Model Combination (Gales and Young, 1993) and HMM decomposition (Varga and Moore, 1990) we do not require models for the acoustic background.

The input to the MS decoder is a set of *coherent source fragments*. Such fragments consist of parameters such as energy estimates in some arbitrary time-frequency region. These fragments have been marked as belonging to a single source by an earlier bottom-up or ‘primitive’ process (Bregman, 1990). Due to speech energy dynamics, it is feasible to find regions with favourable local SNR even if the global SNR is low. Typical primitive processes are those studied in Auditory Scene Analysis: forming time-frequency elements into tracks, or tracks into groups of harmonics with a common fundamental, or exploiting common onset or location in space. In this study we use extremely simple primitive processing, described further in section 7.

MS decoding relies on the missing data recognition techniques developed in earlier sections: each source fragment must be considered as either part of the speech source, in which case the regions it does not contain are interpreted as missing, or belonging to some other source, in which case the fragment itself is missing data.

The result of a multi-source decoding is to establish both the most likely word sequence and also the optimal present/missing labelling. In Bregman’s terms, the MS decoder implements schema-driven grouping (Bregman, 1990).

6.1 Decoding evidence fragments

A brute-force approach to recognising speech from a set of evidence fragments is to evaluate every possible combination of fragments over an entire utterance. Unfortunately, there are 2^N subsets of N fragments, and N could typically become rather large. An alternative approach is to merge decoder hypotheses every time a fragment ends. The complexity then reduces to 2^M where M is the maximum number of *simultaneous* fragments. This is tractable if primitive processes deliver evidence fragments above some minimum granularity, say over some tens of milliseconds duration. Crucially, although N increases with utterance length, M remains essentially constant. Based on the examples in section 7, N can exceed 40 even for short utterances, while M rarely exceeds 6.

The resulting decoder is based on the standard token-passing Viterbi algorithm with the following modifications:

- Tokens keep a record of the fragment assignments they have made i.e. each token stores its labelling of each fragment encountered as either *speech* or *background*.
- When a new fragment starts all existing tokens are duplicated. In one copy the new fragment is labelled as speech and in the other it is labelled as background.
- When a fragment ends we compare, for each state, pairs of tokens that differ only in the label of the fragment that is ending. The less likely token is deleted.
- At each time frame tokens propagate through the HMM as usual. However, each state can hold as many tokens as there are different labellings of the currently active fragments. When tokens enter a state, only those with the same labelling of current active fragments are directly compared. The token with the highest likelihood score survives and the others are deleted.

The scheme may also be seen as a parallel set of normal Viterbi decoders (i.e. with one token for each state) but when a new fragment starts each decoder is duplicated, and when a fragment ends pairs of decoders are merged.

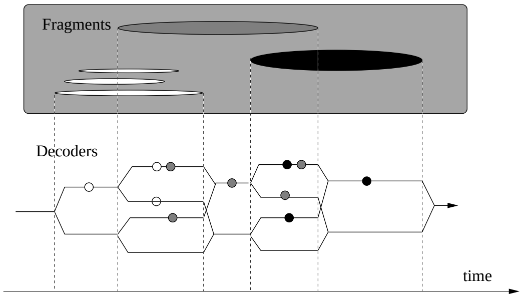

Figure 5: The evolution of a set of parallel decoders. Each parallel path represents a separate decoder, and shaded dots indicate which ongoing fragments are considered as speech. Note that there are at most 4 active decoders, not the 8 required to decode every possible subset of 3 fragments.

6.2 The merging problem

When a fragment ends and decoders are merged, tokens from each decoder are paired up and their likelihoods are compared. However, there is a problem inherent in this comparison: these tokens have arisen from decoders with different speech/background labellings, and as such are calculating missing-data fits based on different patterns of present and missing data. The missing data framework treats the two types of data somewhat differently: the match score for present data is the likelihood calculated by marginalising the full model probability density function over the missing features. However, for the missing data we calculate the 'bounded' probability of the speech being less energetic than the observed background - a true probability rather than a likelihood, and as such not directly comparable. This difficulty has been overcome in previous missing data work, where the amount of present and missing data is the same for each competing hypothesis, by the simple expedient of a scaling factor. However, when comparing decoder hypotheses with differing foreground/background interpretations, a better solution is required.

As a pragmatic solution, the evaluation systems presented in section 7 adopt a scaling of the missing data probabilities by dividing them by the integration range over which they were computed e.g. if the observed value of the background is X, then the speech energy is assumed to lie between 0 and X, and the model probability is computed by integrating the pdf between 0 and X. An 'average likelihood' (i.e. the average value of the pdf over the integration range) is formed by dividing the probability by X, something comparable (when scaled by a fixed constant) with the likelihood values constituting the match core for present data.Further discussion of this issue along with a more principled solution are presented in (Barker et al., 2000a).

7 Multisource Decoder Experiments

The MS decoder architecture is again tested on the connected digit recognition task.

The experiments we report evaluate the new decoding algorithm while using a naive technique to perform the dissection of the spectrogram. This establishes a baseline against which to compare future work that will employ more principled auditory scene analysis techniques.

After deriving the noise estimate mask following the procedure in section 4, the present data mask is dissected by first dividing it into four frequency bands and then labelling contiguous regions within each subband as the separate fragments. This set of fragments and the noisy speech representation are then passed to the MS decoder. Spectro-temporal regions that are not contained in any fragment are assigned a fixed background label.

If the actual noise is non-stationary the noise spectrum estimates, and hence the local SNR estimates, are often grossly inaccurate. A local peak in noise energy can lead to a spectro-temporal region that is mistakenly labelled as having high local SNR. This error then generates a spurious region in the present data mask, usually causing poor recognition. In the new approach, the MS decoder should reject these fragments and label them as background, thereby producing a better recognition hypothesis. This effect is illustrated in Figure 7, where broad-band noise bursts have been artificially added to the noisy data representation. These unexpected components appear as bands in the present data mask and hence disrupt the standard missing data recognition technique ("1159" is recognised as "81o85898"). The third image in the figure shows how the mask is now dissected before being passed into the MS decoder. The final panel shows a *backtrace* of the fragments that the MS decoder marks as present in the winning hypothesis. We see that the noise pulse fragments have been dropped (i.e. relabelled as "background"). Recognition performance is now much improved ("1159" is recognised as "61159").

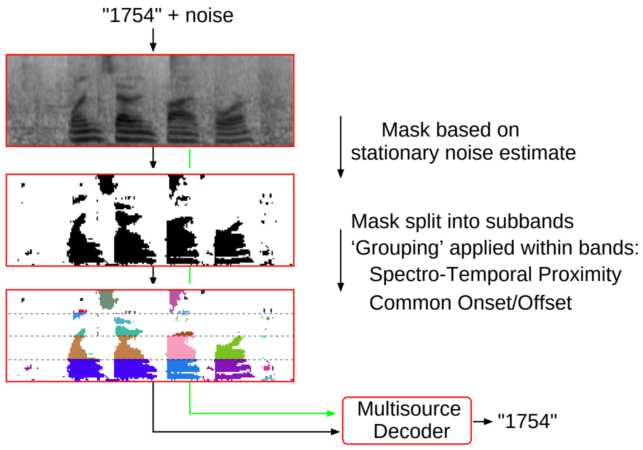
"1754" + noise
Mask based on stationary noise estimate
Mask split into subbands
'Grouping' applied within bands:
Spectro-Temporal Proximity
Common Onset/Offset
Multisource Decoder
"1754"

Figure 6: An overview of the multi-source recognition system.

An initial multisource decoder experiment was run on the same connected digit recognition task of earlier sections, except that a 24 channel filter bank representation was used rather than 32 channels, and decisions were hard not soft. The results in figure 8 compare standard missing data with the new multi-source technique. The scaling constant required to balance missing and present data scores as described in 6.2 was optimally tuned, but a single value sufficed for all noise levels.

It can be seen that multi-source decoding provides a significant improvement at the lower SNRs, e.g. at 5db recognition accuracy is improved from 70.1% to 78.1% – a word-error rate reduction from 29.9% to 21.9%, or 26.7% relative.

One possible cause of the remaining performance gap compared to a priori results is that the fragments supplied to the multi-source decoder are not sufficiently

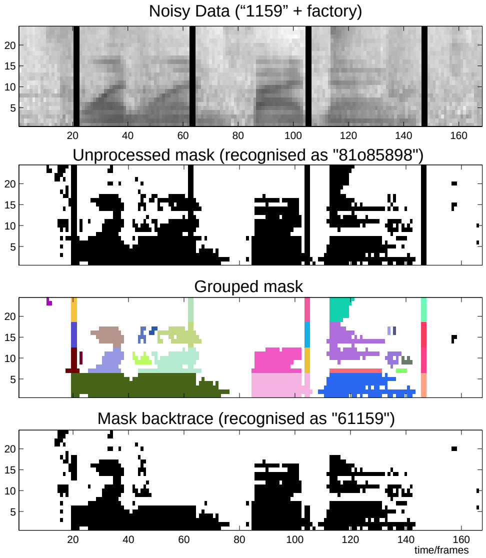

Figure 7: An example of the multi-source system performance when applied to data corrupted by artificial transients (see text).

coherent. In this work we have used a simple set of fragments generated by clumping high energy regions in the SNR mask. If the noise and speech sources occupy adjoining spectro temporal regions this technique will not be able to separate them. This is evident in figure 7 where, as a result of both noise and speech being mixed in the same fragment, a lot of clean speech energy has been removed from the masks and some of the noise energy has survived.

8 Conclusion

Most work on robust speech recognition has been based on the idea of 'reducing the mismatch' between training and test conditions. This leads to the use of noise models and their deployment in techniques such as Spectral Subtraction (Lockwood and Boudy, 1991), HMM decomposition (Varga and Moore, 1990) and Parallel Model Combination (Gales and Young, 1993). If some predictable noise source is present, then it is appropriate to use its statistics in this way, and we have indeed made use of simple noise models in defining our 'missing data masks'. However, if speech is to be recognised within an arbitrary 'auditory scene', such as in a street or at a meeting, the sound sources will not be pre-determined, and they will change in location and with time. For this general case, missing data coupled with multisource decoding has a number of attractions:

- The first stage, the identification of reliable evidence fragments, can be based on processing which exploits only the predictable noise components and low-level constraints such as harmonicity and common onset/offset. There is no need, for instance, to make any assumption about how many sources are present.

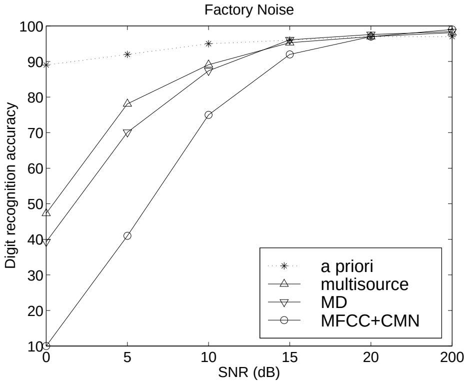

Figure 8: Recognition results for a baseline MFCC system, a missing data system, and the multi-source system.

- The reliable evidence decision does not have to be right all the time, because it can be expressed probabilistically rather than in a discrete way.

- Multisource decoding provides a way in which primitive processing can interact with schema-driven processing, so that the initial grouping stage does not have the responsibility of deciding what belongs to what source.

The scheme we have presented is, arguably, both a competitive robust ASR system and a computational implementation of a psycho-acoustic model.

Much work remains to be done: in the medium terms we plan to

- Use computational auditory scene analysis algorithms to identify evidence fragments.

- Couple the grouping provided by these algorithms with noise estimates to deal with unpredictable sources.

- Develop the more principled formulation of the multisource merging problem outlined in (Barker et al., 2000a).

Acknowledgements

This research is supported by the EC Transfer and Mobility of Researchers network SPHEAR and the EC ESPRIT long term research project RESPITE. L. Josifovski's doctoral studies are supported by Motorola. Thanks to Dan Ellis for helpful discussions about the multisource decoder.

References

Ahmad, S. and Tresp, V. (1993) Some solutions to the missing feature problem in vision. In J. H. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in NIPS 5*, 393–400. Morgan Kaufmann, San Mateo, CA.

Barker, J.P., Cooke, M.P. and Ellis, D.P.W. (2000a) Decoding speech in the presence of other sound sources. In *Proc. ICSLP '00*, IV, 270–273, Beijing, China.

Barker, J.P., Josifovski, L., Cooke, M.P. and Green, P. (2000b). Soft decisions in missing data techniques for robust automatic speech recognition. In *Proc. ICSLP '00*, I, 373–376, Beijing, China.

Bregman, A.S. (1990). *Auditory Scene Analysis.* MIT Press.

Cooke, M.P., Green, P., Josifovski, L. and Vizinho, A. (1999). Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication.* Accepted for publication.

Cooke, M.P. (1991). *Modelling auditory processing and organisation.* PhD thesis, Department of Computer Science, University of Sheffield.

Furui, S. (1997). Recent advances in robust speech recognition. In *Robust speech recognition using unknown communication channels*, 11–20. ESCA-NATO Tutorial and Research Workshop.

Gales, M.J.F. and Young, S.J. (1993). HMM recognition in noise using parallel model combination. In *Eurospeech'93*, II, 837–840.

Ghahramani, Z. and Jordan, M.I. (1994). Supervised learning from incomplete data via an em approach. In J. D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in NIPS 6*, 120–129. Morgan Kaufmann, San Mateo, CA.

Josifovski, L., Cooke, M.P., Green, P. and Vizinho, A. (1999). State based imputation of missing data for robust speech recognition and speech enhancement. In *Eurospeech'99*, VI, 2837–2840.

Lippmann, R. (1996). Speech perception by humans and machines. In *ESCA Workshop on the Auditory Basis of Speech Perception*, 309–316.

Lockwood, P. and Boudy, J. (1991). Experiments with a non-linear spectral subtractor (NSS) Hidden Markov Models and the projection, for robust speech recognition in cars. In *Eurospeech'91*, I, 79–82.

Moore, B.C.J. (1997). *An Introduction to the Psychology of Hearing.* Academic Press, 24/28 Oval Road, London NW1, 4th edition.

Pearce, D. and Hirsch, H.–G. (2000). The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. ICSLP '00*, IV, 29–32, Beijing, China.

Raj, B., Seltzer, M. and Stern, R. (2000). Reconstruction of damaged spectrographic features for robust speech recognition. In *Proc. ICSLP '00*, I, 357–360, Beijing, China.

Renevey, P. and Drygajlo, A. (2000). Introduction of a reliability measure in missing data approach for robust speech recogntion. In *Proc. EUSPICO'2000.*

Steeneken, H.J.M. (1992). *On measuring and predicting speech intelligibility.* PhD thesis, University of Amsterdam.

Strange, W., Jenkins, J.J. and Johnson, T.L. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, 74(3):695–705.

Varga, A.P. and Moore, R.K. (1990). Hidden Markov model decomposition of speech and noise. In *ICASSP'90*, 845–848.

Varga, A.P., Steeneken, H.J.M, Tomlinson, M. and Jones, D. (1992). The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical report, Speech Research Unit, Defence Research Agency, Malvern, U.K.

Vizinho, A., Green, P.D, Cooke, M.P. and Josifovski, L. (1999). Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: An integrated study. In *Proceedings of EuroSpeech'99*, 2407–2410, Budapest.

Warren, R.M., Riener, K.R., Bashford, J.A. and Brubaker, B.S. (1995). Spectral redundancy: Intelligibility of sentrences heard through narrow spectral tilts. *Perception and Psychophysics*, 57(2):175–182.

Young, S.J. and Woodland, P.C. (1993). *HTK Version 1.5: User, reference and programmer manual.* CUED, Speech Group.